Causal Discovery from Discrete Data using Hidden Compact Representation (NIPS 2018)

Ruichu Cai[1], Jie Qiao[1], Kun Zhang[2], Zhenjie Zhang[3], Zhifeng Hao[1,4]

[1]School of Computer Science, Guangdong University of Technology, China

[2]Department of philosophy, Carnegie Mellon University

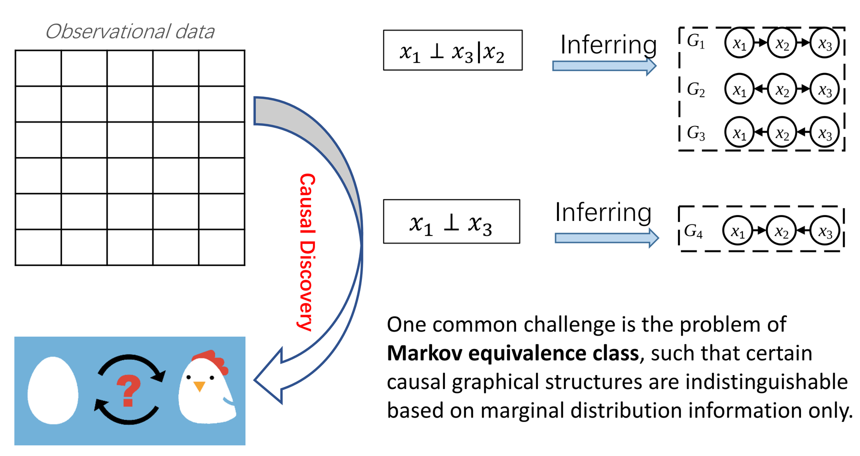
[3]Singapore R&D, Yitu Technology Ltd.

[4]School of Mathematics and Big Data, Foshan University, China

Motivation to Hidden Compact Representation for Causal Discovery

Causal Discovery from Observational Data

Because randomized controlled experiments are usually infeasible and generally too expensive, **observational data-based causal discovery**, has been a focus of recent research.

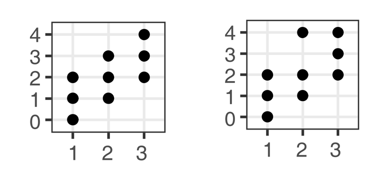


- Challenge
  - Categorical data

Problem and Challenges

Additive Noise Model

$$Y = g(X) + E, X \perp E$$



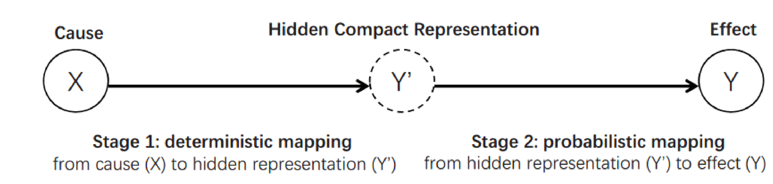
In order to perform an additive operation, ANM needs to encode each category into a set of integers.

After the encoding, the order of each category must consistent to the original one.

- Challenge:
  - It is usually hard to justify the additive operation for discrete data.

Motivations

Food Poisoning: A Hidden Compact Representation Example in Real World.



Stage 1: deterministic mapping from cause (X) to hidden representation (Y')

X \ Y'	Poisonous	Not Poisonous
Poisonous Mushroom	1	0
Mushroom	0	1
Rice	0	1
Poisonous Fish	1	0

Stage 2: probabilistic mapping from hidden representation (Y') to effect (Y)

Y' \ Y	Food Poisoning	Stomach Flu	Normal
Poisonous	0.85	0.10	0.05
Not Poisonous	0.03	0.07	0.90

- First stage
  - Maps cause to a hidden variable of a lower cardinality
- Second stage
  - Generates the effect from the hidden representation.

Hidden Compact Representation Model

Framework

A practical framework for causal inference

Step 1

Estimate the model $M: X \to Y' \to Y$ and $\widetilde{M}: Y \to X' \to X$ by maximizing $\mathcal{L}^*(M;D), \mathcal{L}^*(\widetilde{M};D)$ respectively.

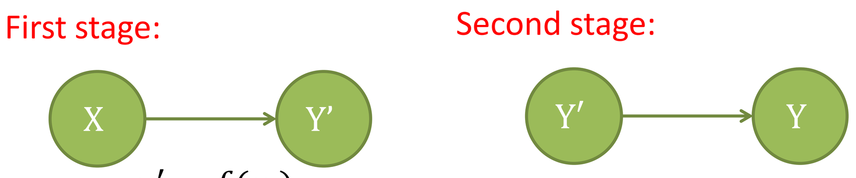
Step 2

If $\mathcal{L}^*(M;D) > \mathcal{L}^*(\widetilde{M};D)$, infer "$X \to Y$"

If $\mathcal{L}^*(M;D) < \mathcal{L}^*(\widetilde{M};D)$, infer "$X \leftarrow Y$"

If $\mathcal{L}^*(M;D) = \mathcal{L}^*(\widetilde{M};D)$, infer "non-identifiable"

HCR



First stage: $y' = f(x_i)$

Second stage: $P(Y = y_i | Y' = f(x_i))$

Given a group of observations $D = \{(x_i, y_i)\}_{i=1}^m$, the log-likelihood of the model $M: X \to Y' \to Y$:

$$\mathcal{L}(M;D) = \log \prod_{i=1}^{m} \sum_{y_i'} P(X = x_i) P(Y = y_i | Y' = f(x_i))$$

$$= \sum_x n_x \log\left(\frac{n_x}{\sum_x n_x}\right) + \sum_{y'} \sum_y n_{y',y} \log\left(\frac{n_{y',y}}{\sum_y n_{y',y}}\right)$$

To select the best causal model M, we regard the model with the highest $\mathcal{L}^*$ (BIC) as the most suitable one.

$$\mathcal{L}^*(M;D) = \mathcal{L}(M;D) - \frac{d}{2}\log(m)$$

where $d = (|X| - 1) + |Y'| (|Y| - 1)$ measures the effective number of parameters in the model.

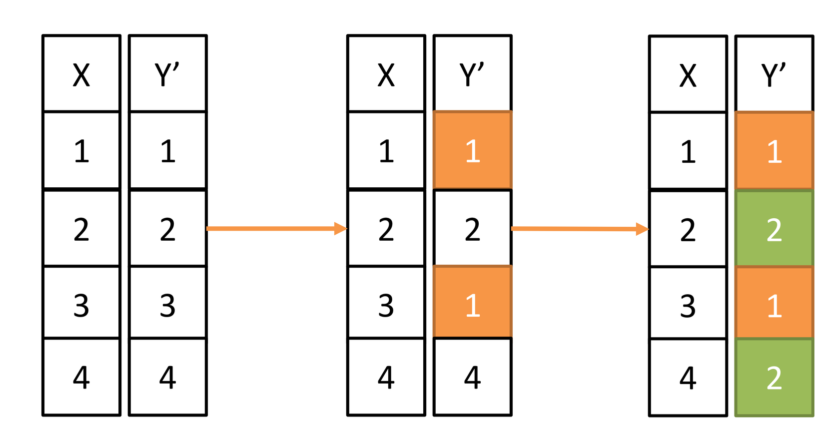
Optimization

Technical details addressing the optimization scheme with hidden compact representation.

An alternate maximization procedure: $\max \mathcal{L}^* = \sup_f \max_\theta \mathcal{L}^*$

i) Estimate the MLE of $P(X)$ and $P(Y|Y')$, while fixing the function f: $\hat{\theta} = \underset{\theta}{argmax}\, \mathcal{L}(\theta; D)$

ii) Choosing the best f such that: $f = \sup_f \mathcal{L}^*(f; D)$



Identifiability

Asymptotic Correctness

We shall show that under the hidden compact representation model, the causal direction is identifiable in the general case (under some technical conditions)

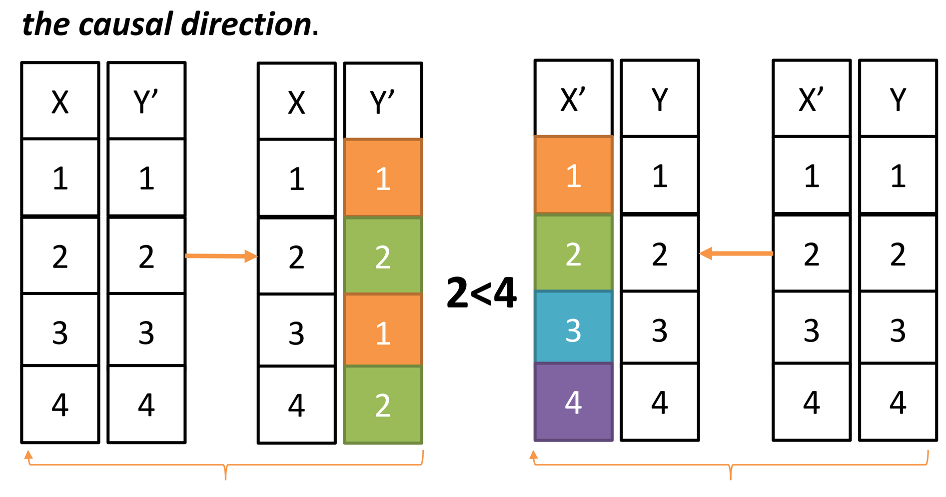
Theorem 1. *Assume that for the correct causal direction, the conditional distribution $P(Y | X)$ is random in the sense that*

A1. there does not exist values $y_1 \neq y_2$ such that $P(Y = y_1 | X)$ equals $P(Y = y_2 | X)$ times a constant for all possible X values. (Note that both $P(Y = y_1 | X)$ and $P(Y = y_2 | X)$ are functions of X.)

Then asymptotically, in the reverse direction there does not exist $X' = \hat{f}(Y)$ with $|X'| < |Y|$ such that $P(X | Y) = P(X | X')$ for all possible X and Y values, i.e., the reverse direction does not admit a low-cardinality hidden representation $\hat{f}(Y)$.

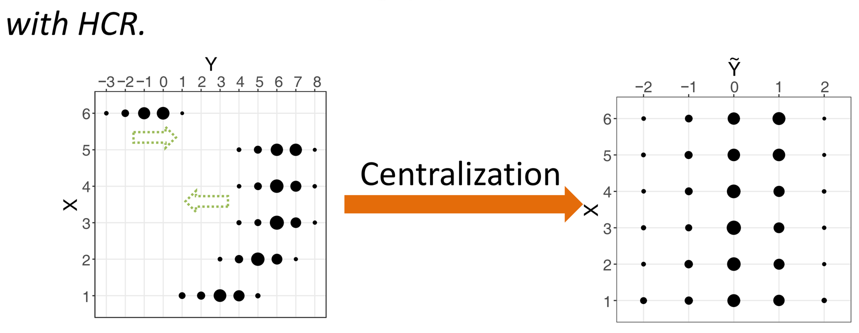
Identifiability

Theorem 2. *Assume that in the causal direction there exists the transformation $Y' = f(X)$ such that $P(Y | X) = P(Y | Y')$, where $|Y'| < |X|$, and assumption A1 holds. Then to produce the same distribution $P(X, Y)$,* ***the reverse direction must involve more effective number of parameters in the model than the causal direction****.*



Generalization

Given a cause-effect pair $X \to Y$ following the discrete additive noise model ($Y = g(X) + E, X \perp\!\!\!\perp E$), after preprocessing each sample using $\hat{Y} := Y - g(X)$, its causal direction is identifiable with HCR.



Intuitively, the data generated from the discrete additive noise model can be compacted into in a hidden compact representation with $|Y'| = 1$, after a preprocessing for each sample using $\hat{Y} := Y - g(X)$

Experiments

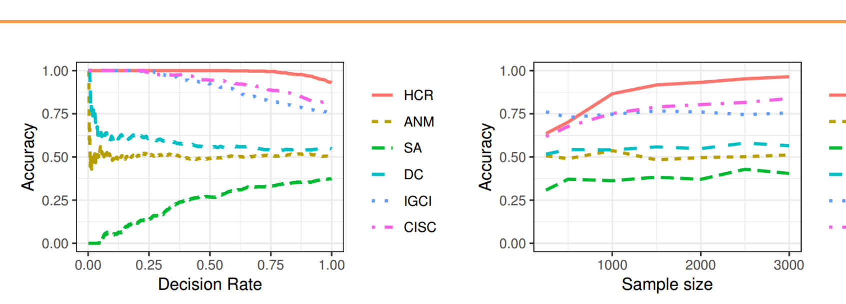
Synthetic Data

Experiment setting

In all the experiments, we generate 1000 different causal pairs and 2000 samples for each pair.
The HCR data are generated according to the following two-stage procedure.

1. Sample X according: $\mathcal{X} \sim U\{3, 4, \dots, 15\}$, then randomly assign Y' from $\{1, 2, \dots, |\mathcal{X}|\}$.
2. sample Y according to Y' and $P(Y | Y')$, and $|Y|$ is generated from the interval $\{|Y'|, \dots, 15\}$

HCR data



(a) Sensitivity to Decision Rate. (b) Sensitivity to Sample Size.

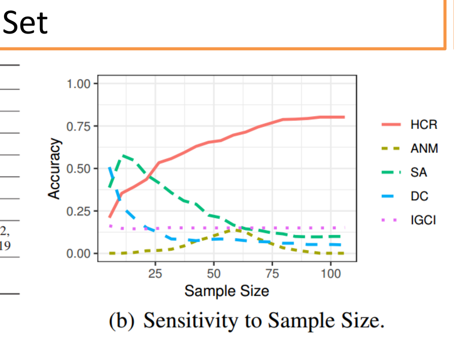
Real-World Data

Experiment on real-world datasets

Results on Pittsburgh Bridges Data Set

Ground truth	X	Y'	Y
Erected→Span	Crafts	1	Medium: 0.5, Short:0.5
	Emerging,Mature, Modern	2	Long: 0.37, Medium:0.59, Short:0.04
Material→Span	Steel	1	Long: 0.42, Medium:0.58
	Iron,Wood	2	Medium:0.55, Short:0.45
Material→Lanes	Steel	1	2 Lane:0.6, 4 Lane:0.33, 6 Lane:0.06
	Iron,Wood	2	1 Lane:0.15, 2 Lane:0.8, 4 Lane:0.04
Purpose→Type	Aqueduct, Highway, Walk	1	Arch:0.18, Cantilev:0.12, CONT-T:0.12, Simple-T:0.24, Suspen: 0.15, Wood:0.19
	RR	2	Cantilev:0.06, CONT-T:0.03, Simple-T:0.81, NIL:0.3,wood:0.06

(a) Hidden Compact Representation.



(b) Sensitivity to Sample Size.

Analysis

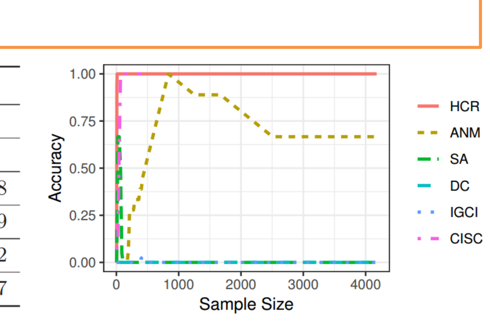
Erected → Span:
Crafts is the main cause of the medium and short bridge

Material → Span and *Material → Lanes:*
The steel belongs to modern material with high strength, while iron and wood are classic materials with lower strength.

Results on Abalone Data Set

Abalone	X	Y'	Y
Sex→Length	Infant	1	0.43 ± 0.1
	Female, Male	2	0.57 ± 0.96
Sex→Diameter	Infant	1	0.33 ± 0.088
	Female, Male	2	0.45 ± 0.079
Sex→Height	Infant	1	0.11 ± 0.032
	Female, Male	2	0.15 ± 0.037

(a) Hidden Compact Representation



(b) Sensitivity to Sample Size.

Analysis

Sex → {Length, Diameter, Height}:
Y' indicate the childhood and adulthood of the abalone in the real world. In the second stage, the mapping shows the maturity causes the sizes